LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-PROC-209783

Inverse Scattering Approach to Improving Pattern Recognition

G. Chapline, C-Y. Fu

February 16, 2005

SPIE Defense and Homeland Security Symposium
Orlando, FL, United States
March 28, 2005 through April 1, 2005

Inverse Scattering Approach to Improving Pattern Recognition

George Chapline and Chi-Yung Fu
Lawrence Livermore National Laboratory, PO Box 808, Livermore, CA 94551

Abstract

The Helmholtz machine provides what may be the best existing model for how the mammalian brain recognizes patterns. Based on the observation that the "wake-sleep" algorithm for training a Helmholtz machine is similar to the problem of finding the potential for a multi-channel Schrodinger equation, we propose that the construction of a Schrodinger potential using inverse scattering methods can serve as a model for how the mammalian brain learns to extract essential information from sensory data. In particular, inverse scattering theory provides a conceptual framework for imagining how one might use EEG and MEG observations of brain-waves together with sensory feedback to improve human learning and pattern recognition. Longer term, implementation of inverse scattering algorithms on a digital or optical computer could be a step towards mimicking the seamless information fusion of the mammalian brain.

Keywords : Helmholtz machine, inverse scattering, brain-waves

1. Introduction

It has been understood for some time that pattern recognition systems are in essence machines that utilize either preconceived probability distributions or empirically determined posterior probabilities to classify patterns [1]. In the ideal case where the a priori probability distribution $p(\alpha)$ for the occurrence of various classes α of feature vectors and probability densities $p(x|\alpha)$ for the distribution of data sets x within each class are known, then the best possible classification procedure would be to simply choose the class α for which the posterior probability

$$P(\alpha \mid x) = \frac{p(\alpha)p(x \mid \alpha)}{\sum_{\alpha} p(\alpha)p(x \mid \alpha)} \tag{1}$$

is largest. Unfortunately in the real world one is typically faced with the situation that neither the class probabilities $p(\alpha)$ nor class densities $p(x|\alpha)$ are precisely known, so that one must rely on empirical data to estimate the conditional probabilities $P(\alpha|x)$ needed to classify data sets. In practice this means that one must adopt a parametric model for the class probabilities and densities, and then use empirical data to fix the parameters θ of the probability model. Once values for the model parameters have been fixed, then input patterns can be classified by substituting the values for the probabilities $p(\alpha; \theta)$ and $p(x|\alpha; \theta)$ into equation (1), and then choosing the class α which maximizes the conditional probability $P(\alpha|x)$.

Unfortunately determining values for the model parameters from empirical data is itself a computationally intractable problem. This means that in practice one is usually limited to using models of relatively modest complexity, and consequently one is always faced with the issue of choosing the best possible values for the model parameters. One popular way of measuring how good a particular set of model parameters is at reproducing the observed data, known as the maximum likelihood (ML) estimator [2], can be motivated by noting that the formula for the posterior probability given in equation (1) can be formally interpreted as the canonical Boltzmann distribution for the population of energy levels of a physical system in equilibrium with a heat bath. In particular if one defines the "energy" of a classification α to be

$$E_{\alpha} = -\log p(\alpha)p(x \mid \alpha) \tag{2}$$

then the posterior probability introduced in (1) can be formally expressed in the form

$$P(\alpha \mid x) = \frac{e^{-E_\alpha}}{\sum_\alpha e^{-E_\alpha}} \tag{3}$$

where the energies are those defined in equation (2). If we imagine that the "energy levels" E_α defined in equation (2) define a fictitious physical system whose levels are populated according to the canonical Boltzmann distribution (3), then the thermodynamic free energy of this system will be given by

$$F(x) = \sum_\alpha \{E_\alpha P(\alpha) + (kTP(\alpha)\log P(\alpha))\}, \tag{4}$$

where we have used $P(\alpha)$ as shorthand notation for the canonical distribution (3). If instead of the true probability distributions $p(\alpha)$ and $p(x|\alpha)$ one uses model probabilities $p(\alpha; \beta)$ and $p(x|\alpha; \beta)$ to calculate a tentative probability distribution $P_\beta(\alpha)$ for different classifications of an input data set x, then equation (3) will no longer necessarily be satisfied and the free energy calculated from equation (4) will in general differ from the true free energy. In particular we would have that

$$F(x) = F(x, \beta) + kT\sum_\alpha P_\beta(\alpha)\log[P_\beta(\alpha) / P(\alpha)] \tag{5}$$

where $F(x,\beta)$ is the free energy calculated using the estimated distribution $P_\beta(\alpha)$. The quantity $\sum_\alpha P_\beta(\alpha)\log[P_\beta(\alpha) / P(\alpha)]$ in the second term in equation (5) is always positive and measures of the difference in bits between the model distribution $P_\beta(\alpha)$ and the true distribution $P(\alpha)$. This distance measure, known as the *Kullback- Leibler divergence,* is the basis for maximum likelihood estimates. It should be noted that the estimated free energy $F(x, \beta)$ is always greater than the true equilibrium free energy $F(x)$, so that a good choice for the estimated probability distribution is one that which minimizes the estimated free energy $F(x, \beta)$.

In the following we will focus on a particular method for constructing model probability distributions that minimize the estimated free energy known as the *Helmholtz machine*. The name is inspired by Helmholtz's suggestion that the mammalian brain functions as a statistical inference engine. In the work of Hinton et al [3,4] a "wake-sleep" training algorithm was used to fix the parameters of the Helmholtz machine. Our new suggestion, motivated by the similarity between the wake-sleep algorithm and the problem of determining the potential for a multi-channel Schrodinger equation from scattering data, is that inverse scattering techniques might be used both to improve human cognition and also provide new computational strategies for information fusion.

2. Bayesian networks and Helmholtz machines

In a Bayesian network [5] causal relationships between pattern features are represented by a probability distribution for the states of the nodes in a layered network, which is a product of conditional probabilities for each unit given the values of the units which proceed it in some ordering. Many types of Bayesian networks are possible, but typically such networks have a tree-like structure and the conditional probabilities have the Markov property; i.e. the random variables at nodes not connected by a branch are conditionally independent given those variables which are so connected. It is the Markov structure of Bayesian decision trees, which provides an entry for utilizing the formalism of quantum mechanics to solve pattern recognition problems.

One of the main practical problems with using Bayesian networks to search for patterns is that, given a decision tree and associated set of conditional probabilities which constitute a model for the world, finding explanations for input data will typically involve using Markov chain Monte

Carlo methods to invert the world model. Because of the need for repetitive sampling, this cannot in general be done in real time. Remarkably, though, the Helmholtz machine introduced by Hinton et.al. [3,4] offers the promise of alleviating these problems.

The Helmholtz machine is a layered network of units whose activation levels are quantized to be either 0 or 1. Some of the binary units represent environmental input data, while the remaining hidden units represent possible explanations for the input data. All information concerning conditional probabilities is contained in the values of the connection strengths w_{ij} between nodes; indeed it is possible that these weights play much the same role in the Helmholtz machine as the synaptic connections in the cerebral cortex. In the scheme of Hinton et al the activation of the i^{th} hidden unit in the n+1 layer is chosen stochastically in accordance with the probability $p_j(x)$ given by

$$p(a_i(n+1) \mid a(n)) = \sigma[\sigma(1-2a_i(n+1))\sum_j w_{ij}a_j], \tag{6}$$

where $\sigma(x) = 1/[1+\exp(-x)]$. The vector $a(n) = \{a_i(n)\}$ in equation (6) denotes the set of activation levels at layer n of the network. If one assumes that the activities of the binary units within a given layer are assumed to be independent, then the probability of a particular explanation $\alpha = \{a(n), n>1\}$ for the input data will be given by the product :

$$Q(\alpha) = \prod_{n>1} \prod_j [p(a_j)]^{a_j}[1 - p(a_j)]^{1-a_j} \tag{7}$$

All of the information concerning the structure of the external world that is needed to explain a given set of input patterns is encoded into the connection weights w_{ij}, and the main obstacle to implementing the just described recognition network is determining the values of these parameters.

In the "wake-sleep" training scheme of Hinton et. al. the the connection weights and biases of a separate "top-down" network are used offline to generate a "true" posterior probability distribution $P_\alpha(\beta)$. The connection strengths w_{ij} which are used in the recognition network to explain input patterns are determined simultaneously with the parameters of the top-down model by using gradient descent methods to minimize the estimated free energy function F(x,Q). The "bottom-up" conditioning of virtually every kind of neural network that has be used to interpret practical sensor data is not only be computationally tedious, but also problematic from the point of view as to whether the network is really giving the correct weightings to various possible explanations. The introduction of a separate "top-down" network for this purpose by Hinton et. al. appears to be a significant step towards both reducing the computational complexity of pattern recognition; particularly when there are many alternative explanations for the input data.

3. Schrodinger representation for a Helmholtz machine

As noted above the transition probabilities which describe the evolution from one layer to the next in either the bottom-up or top-down networks of a Helmholtz machine satisfy the identities required for Markov probabilities. Although Markov processes are usually thought of as progressing forward in time, any Markov chain can be run in reverse and there is in fact a time symmetric or "Schrodinger" representation for Markov chains [6]. This time symmetric representation allows one to express the transition probabilities for either the forward or time reversed Markov chain in terms of real quantities p^+ and p^- which replace the amplitude A and phase θ of a Schrodinger wavefunction as follows:

$$p^+ \equiv A\exp(\theta)\ ,\ p^- \equiv A\exp(-\theta) \tag{8}$$

The forward and backward Markov transition probabilities can then be expressed in the form:

$$\begin{aligned} P^{\uparrow}[b(n+1) \mid a(n)] &= [p^+(a,n)]^{-1}G(a,n\,;\,b,n+1)\,p^+(b,n+1), \\ P^{\downarrow}[b(n+1) \mid a(n)] &= p^-(a,n)G(a,n\,;\,b,n+1)\,[p^-(b,n+1)]^{-1}, \end{aligned} \tag{9}$$

where G is the Green's function for the multi-channel imaginary time Schrodinger equation.

In principle equations (9) could be used to simulate the dynamics of any quantum system; although in practice one would probably be limited to quantum systems for which the dimension of Hilbert space is not too large. However, for the purposes of trying to model how the mammalian brain works, it is more useful to think of Eq's (9) as allowing one to use the formalism of quantum mechanics to model the way a Helmholtz machine recognizes patterns. In the case where each node of the Markov network has just two states (viz "spin up" or "spin down"), the network can be identified with Hinton's original version of the Helmholtz machine [3,4], and we see that the Schrodinger representation, Eq's (9), provides a way of representing the connection strengths of the Helmholtz machine in terms of quantum mechanical transition amplitudes. Furthermore, we can now view the problem of determining the connection strengths in a Helmholtz machine as equivalent to the "inverse problem" of determining the potential for a multi-channel Schrodinger equation from knowledge of initial and final wavefunctions for the quantum system. This latter inversion problem has been intensively studied in the particular case of 3-dimensional scattering of waves from a non-spherical potential [7,8], and also occurs in the context of adaptive optics [9]. Thus there is an existing body of mathematical knowledge concerning the inversion of scattering data that one can take over for the purpose of modeling how the mammalian brain extracts information from sensory inputs.

4. Adaptive optics model

The discussion of section 3 suggests that inverse scattering techniques applicable to multi-channel quantum mechanics may lead to new approaches to pattern recognition. As it happens the inversion problem for multi-channel quantum mechanics has previously made an appearance in an engineering problem that may provide valuable insight into how the mammalian brain fuses different kinds of sensory data. This problem is adaptive optics in the presence of photon noise. The connection of this problem with the inversion problem for multi-channel quantum mechanics was first pointed out by Freeman Dyson [9].

The system considered by Dyson is a deformable reflecting mirror, where the shape of the surface is adjusted so as to just compensate for small shifts $\Delta(\xi,t)$ in the optical path length of light rays incident on the surface at various locations ξ. Dyson proceeds by writing down equations describing the interplay between small deformations in the shape of a surface and changes in the intensity of light on an array of sensors which measures the shape of the wave front. The first equation supposes that we have a control system that adjusts the displacement $\Delta(\xi,t)$ of the surface with sufficient accuracy so that the intensity of light on the sensor array at time t and position x is a linear function of the error $e(\xi,t) \equiv \Delta(\xi,t) + \Delta(\xi,t)$:

$$I(x,t) = I_0(x) + \int d^2\xi \; B(x,\xi)e(\xi,t), \tag{10}$$

where I(x,t) is the recorded intensity of light on the sensor array at time t and position x, and $I_0(x)$ is the recorded intensity on the sensor array when the surface ξ is illuminated in the absence of imposed variations in phase with respect to position or time. The second equation relates the deformation $\Delta(\xi,t)$ to the intensity of light on the phase measuring array:

$$\Delta(\xi,t) = \int d\xi \int_{-\infty}^{t} dt' \; A(\xi,x,t') \, I(x,t), \tag{11}$$

where the integral over $d\xi$ means sampling the light intensity at a sufficiently large number of points on the sensor array as is required to determine the parameters ξ_1, ξ_2,... which define the shape of the surface. When photon noise is neglected, equations (10) and (11) have the classical solution (in matrix shorthand)

$$e = [1 - AB]^{-1} \; . \tag{12}$$

Eq. (12) shows that when the negative feedback is sufficiently strong the error $e(\xi,t)$ can be reduced to a small fraction of the change in optical path length. That is, in the absence of noise one can cause the position of the surface at each point to just track the change in optical path length.

What is perhaps most remarkable about the adaptive optics problem, though, is that when photon noise is taken into account the problem of adjusting the shape of the surface to compensate for changes in the optical path of the illuminating beam becomes equivalent to solving the multi-channel Schrodinger equation. This situation is qualitatively different from the classical case because the negative feedback in Eq. (11) will amplify the photon noise. It is not hard to show that in the presence of noise the two point correlation function for the path length errors e(□,t) averaged over a time long compared to the characteristic time for photon number fluctuations has the form (again in matrix shorthand):

$$< e_1 e_2> = [1-A_1 B_1]^{-1} [1- A_2 B_2]^{-1} \{U_{12} + A_1 A_2 \square_{12} I_0\}, \quad (13)$$

where U_{12} is the average $<\square_1\square_2>$ over the same time and $\square_{12} = \square(\square_1 - \square_2) \square (t_1 - t_2)$. In contrast with the classical case, an optimal choice for the feedback matrix A is somewhat arbitrary. Dyson takes as the criterion for optimizing the feedback system that a quadratic function of the feedback errors should be minimized, in which case the optimal feedback matrix $A(\square,x,t')$ can be expressed in the form: $A = KB^T I_0^{-1}$ where the matrix $K(\square_1 , \square_2 , t_1 - t_2)$ satisfies

$$K + K^T + K(B^T I_0^{-1}B)K^T + U = 0 \ . \quad (14)$$

Eq. (14) is the central equation of inverse scattering theory for the Schrodinger equation [8].

In the context of modeling the mammalian brain the labels representing positions on the deformable mirror would denote specific sets of synaptic connections within the cerebral cortex associated with recognizing specific features of the sensory inputs. The role of surface deformations of a mirror would be played by changes in the strengths in these connections induced by sensory habituation. The plasticity of the human brain to sensory habituation is a well known phenomenon, and could provide a practical basis for using inverse scattering algorithms to model human pattern recognition.

5. Improving human learning and cognition

The idea here is to use magnetoencephalographic (MEG) and/or electroencephalographic (EEG) observations of brain-waves to extract the phase □ as defined in Eq. (8). This phase represents the difference between an explanation chosen from the ensemble of possible explanations that have been stored in memory, and what is actually observed. Thinking of this difference as a "phase" has the advantage that the adaptive optics analogy provides an easily visualized conceptual framework for information processing that might permit us to glimpse for the first time how the mammalian brain seemingly effortlessly extracts subtle patterns from multiple sensor inputs. We note in passing that our quantum mechanics related "phase" is a purely formal construct, and that we are not suggesting that real holographic-like interference plays a role in the human brain, as has been suggested [10].

Actually brain-wave patterns are rather noisy, but previous experience with neural network based algorithms for extracting the phase for different optical paths from measurements of the phase across a transverse surface in an optical system [11] could be very helpful. In adaptive optics systems the phase obtained by "smoothing" interferometric measurements of the phase across a 2-dimensional array is used to determine the surface displacements of a deformable mirror which will cancel out the distortions in the image resulting from variations in optical path length due to atmospheric turbulence. Of course, in the case of the human brain, distortions in EEG and MEG signals may arise not just from noise, but also because the meaning of the sensory inputs may be ambiguous. The development of methods for extracting the phase □ from brain-waves, and then comparing this phase with that expected for likely interpretations of the sensory inputs would allow one to monitor the progress of human pattern recognition and assess the efficiency of various teaching techniques for improving cognition.

One key element to achieving success with such a program is to obtain significant MEG or EEG data in real time without relying on averaging over many observations as is typically done in the field. This is an important step because when averaging over many epochs, the brain will typically adapt to the signals. As a result, in order to study the dynamic of the brain's response, real time signal extraction is essential. Current standard practice is to repeat the stimulation 100 to 300 times to obtain 100 to 300 epochs of brainwave patterns followed by averaging to reduce noise. We have successfully designed a new unsupervised signal extraction algorithm for such

purpose. Preliminary results are encouraging. Figure 1 shows the epoch-by-epoch "cleanup" MEG signal adaptation of the brain as indicated by the decreasing response of the valley found at ~250 msec based on the outputs from a SQUID device to an audio stimulation. Different epochs were offset along the y-axis to show the adaptation of the brain. As a comparison, we also show the averaging result of the 120 epochs of raw MEG signals at the top. A100 fold decrease in experiment time has been achieved and this success is important for the practical implementation of our ideas.

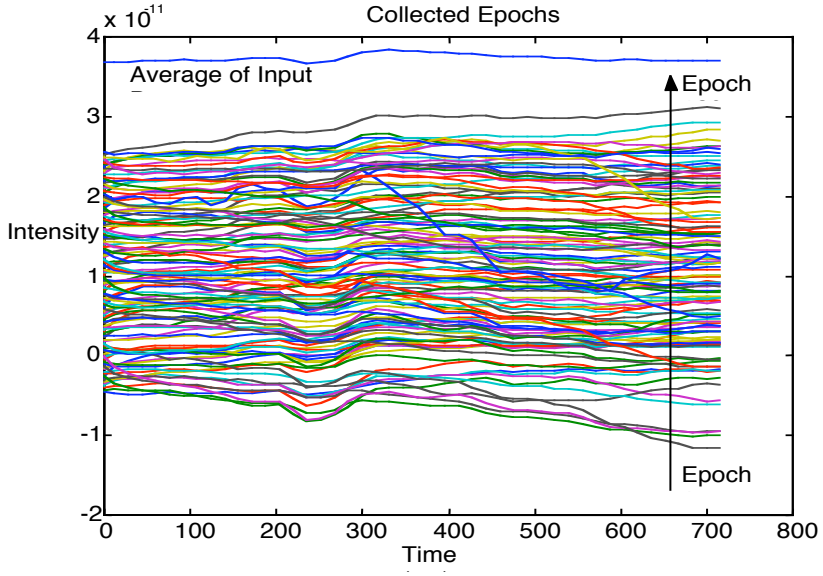

Figure 1 – Real time extraction of MEG data epoch-by-epoch

With "clean" EEG and MEG signals in hand one can attempt to extract the phase differences between the brain-wave signals associated with sensory inputs and the brain-wave correlates for various explanations of the sensory data. It is known, for example, that the meanings of words or phrases have brain-wave correlates [12], and so these expected responses to audio inputs could be compared with the observed brain-wave response. One could then proceed to monitor changes in the phase difference in order to assess the efficacy of various methods for teaching word recognition such as phonological priming [13].

6. Information Fusion

In our proposal to improve human cognition via the application of inverse scattering theory the human brain itself was the "Helmholtz machine". However, one could also contemplate implementing a Helmholtz machine-like model for the mammalian brain directly on a massively parallel computer. For example, one might consider using the independent processors in a massively parallel digital computer as the nodes of a Helmholtz machine network; with time evolution of the nodes representing the passing of information from one layer to the network. During both the "wake" and "sleep" modes of operation each node could update itself according to Eq. (6) using information about the current state of the nodes that is stored in a shared memory. Of course, in pursuing such a program one is faced with the usual training problem afflicting all artificial neural networks; namely, the determination of the connection strengths w_{ij} from training data is computationally tedious when there are many input channels.

As an alternative one might consider using inversion techniques that have been developed for the 3-dimensional wave equation (see e.g. ref. 8). One still needs to develop feedback algorithms for recognizing reference images, but a considerable body of knowledge regarding the inversion problem for the 3-dimensional wave equation could be brought to bear that may make the computational problem of representing the plasticity of the cerebral cortex more tractable. An additional intriguing possibility would be to use optical computing directly. Representing

information via the amplitude and phase of coherent light would make the fusion of different kinds of information seem "seamless"; which is, of course, one of the things that makes the mammalian brain seem so miraculous. Eventually we would like to investigate the relationship between information fusion in Bayesian networks and brain activities bound together by the gamma oscillation [14]. In these cases we expect that an optical representation of sensory feedback affecting multiple sensory modalities will provide new insights into conscious awareness.

7. Summary

We have proposed here a new approach to pattern recognition which is partly inspired by our current understanding of how the mammalian brain recognizes patterns. This approach is based on inverse scattering techniques as used, for example, in adaptive optics systems. We hope to apply our technique to improve real world human pattern recognition and to assess the efficacy of various teaching methods by using neural network techniques to "clean up" observed EEG and MEG measurements, and then compare these preprocessed signals with expected brain-wave correlates. In the guise of defining a potential for a Schrodinger equation this work will bear directly on the fundamental question as to how well Bayesian models of conditional probability explain the architecture and dynamics of the cerebral cortex. For example, the success or failing of inverse scattering techniques should provide considerable insight into how well Bayesian models can represent the effects of training on perception.

Acknowledgements

The authors wish to thank Roger Werne for encouragement and many discussions and Loren Petrach for computational support. This work was performed in part under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

References

1. B. D. Ripley, *Pattern Recognition and Neural Networks* (Cambridge University Press 1996).
2. S. Kullback, *Information Theory and Statistics* (Wiley, 1959).
3. G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal, "The 'Wake- Sleep' Algorithm for Unsupervised Neural Networks", Science, **268**, 1158-116 (1995).
4. P. Dayan, G. E. Hinton, and R. M. Neal, "The Helmholtz Machine",Neural Comp. **7**, 889-90 (1995).
5. J. Pearl, *Probabilistic Inference in Intelligent Systems* (Morgan Kaufmann 1988).
6. M. Nagasawa, *Schrodinger Equations and Diffusion Theory* (Birkhauser, 1993).
7. R. G. Newton, *Inverse Schrodinger Scattering in Three Dimensions* (Springer- Verlag, 1989).
8. M. Cheney and J. Rose, "Generalization of the Fourier Transform: Implications for Inverse Scattering Theory", Phys. Rev. Lett. **60**, 1221-1224 (1988).
9. F. J. Dyson, "Photon noise and atmospheric noise in active optical systems", J. Optical Soc. Am. **65**, 551-558 (1975).
10. K. H. Pribram, *Brain and Perception* (Lawrence Erlbaum, 1991).
11. C. Y. Fu, L.I. Petrich, G. F. Chapline, and S. S. Olivier, "Intelligent Wavefront Reconstructors for Adaptive Optics," to be submitted.
12. P. Suppes, B. Han, and Z-L. Lu, "Brain-wave recognition of sentences", Proc. Natl. Acad. Sci. USA **95**, 15861-15866 (1998).
13. L. Slowiaczek, H. Nusbaum, D. Pisoni, "Phonological Priming in Auditory Word Recognition", J. Expt. Psychology, **13**, 64-75, 1987.
14. O. Bertrand and C. Tallon-Baudry, "Oscillatory gamma activity in humans: a possible role for object representation", Int, J. of Psychophysiology, **38**, 211-223 (2000).